Multi Class Pulmonary Disease Prediction Using Genetic Algorithm Optimization

Priyanka Singh
(Research Scholar)
Computer Science & Engineering
RNTU, Bhopal

Dr S. Veenadhari
(Dean)
CS/IT Department
RNTU, Bhopal

***Abstract*—** Chest X-ray (CXR) imaging is acknowledged as a widely used and cost-effective diagnostic test capable of identifying a broad spectrum of diseases. Challenges in Disease Diagnosis from CXR Samples: Despite its prevalence, diagnosing diseases accurately from CXR samples remains a challenge even for experienced radiologists. Focus on Pulmonary The study specifically targets Pulmonary Disease Prediction, indicating a specialization within the broader medical context. Weiner filtering is introduced as a technique applied to the input image to eliminate noise, enhancing the quality of the image for subsequent analysis. Co-occurrence matrix features are extracted from the processed image, providing a set of statistical measures capturing spatial relationships between pixels. To increase prediction accuracy, the study employs the Biogeographical Optimization algorithm for clustering. Training is conducted on the clustered images to enhance the model's ability to discern patterns and features. The clustered image and co-occurrence matrix features extracted during the process are then utilized for training a neural network. Neural networks are powerful machine learning models capable of learning complex patterns and relationships from data. Experiments are carried out on both binary and multi-class image datasets. The results indicate that the proposed model outperforms existing models, showing improvements in evaluation parameter values. Evaluation parameters could include metrics like accuracy, precision, recall, and F1 score, depending on the specifics of the study.

Index Terms— Chest x-ray images, Image processing, Genetic Algorithm, machine learning.

I. INTRODUCTION

The healthcare industry generates an extensive volume of data related to patients, illnesses, and diagnoses. Unfortunately, this wealth of information often lacks its due significance due to inadequate analysis. Numerous studies have explored the effectiveness of computer-aided diagnoses in collaboration between medical researchers and computer scientists. Some systems in this realm, resembling expert systems, aim to emulate the decision-making processes of medical professionals. Additionally, computer-aided detection systems can handle intricate clinical data, presenting an opportunity to provide valuable insights to clinicians for enhancing diagnostic accuracy [1].

Using smart computers helps doctors a lot in figuring out diseases Machine learning a fancy term for these smart computers is super helpful in figuring out what sickness someone might have putting them in the right category and predicting what might happen next Important tools like special X rays CT scans and chest pictures CXR imaging are crucial in spotting lung problems especially during big health crises like the COVID 19 outbreak [1–4].

Even though scientists have made good progress in finding lung diseases they now want to make these smart computers even

better by using something called edge computing Right now the systems we have for sorting out lung diseases mostly work alone or use the internet cloud to help This study uses a bunch of pictures CXR dataset collected by Rahman [5 6] that anyone can use for research [5] They first used these pictures to find and sort out COVID 19 and viral pneumonia [7]

The smart computers called models use deep neural networks and do a great job figuring out details in pictures to help find diseases They can find tuberculosis sort out different lung diseases in pictures and even spot little growths in special X ray pictures CT scans But here's the challenge there aren't enough super trained doctors to do all this especially to tell COVID 19 apart from other diseases X ray pictures are still a very common way to figure out what's wrong in your chest Now with better machines and computer smarts these computers can learn from a lot of pictures and get as good as human doctors in figuring things out This is a big deal because we don't always have enough doctors and these smart computers can be just as accurate on big sets of pictures.

The global deficit in trained radiologists accentuates the imperative of delving into diverse approaches encompassing machine learning ML deep learning DL and transfer learning TL This study aspires to scrutinize the effectiveness of these varied approaches in disease classification utilizing an openly accessible CXR image dataset

II. Related Work

A cutting-edge method that uses deep learning to automatically identify pneumonia was presented by Gabruseva et al. [8]. Built atop the RetinaNet foundation, their model has an impressive capacity to categorize pictures into three groups: "Normal," "No Lung Opacity/Not Normal," and "Lung Opacity." This model is noteworthy for achieving the highest validation mean average precision (mAP), which is 0.250230. This indicates that the model is reliable and precise in recognizing instances of pneumonia.

Ibrahim et al. [9] presented a sophisticated deep learning method for identifying pneumonia in chest X-ray pictures in a different research. Using the robust AlexNet architecture, their model is evaluated using measures such as overall test accuracy, sensitivity, and specificity, and provides results for many pneumonia categories. These analyses are specifically designed to distinguish between cases of viral pneumonia other than COVID-19 and healthy patients, offering a detailed appraisal of the model's efficacy.

The authors of the paper by Hasan et al. [10] used a creative optimization technique. Four heuristic algorithms were integrated with the Self-Organizing Map (SOM) optimization method. The method's adaptability was demonstrated by using this all-encompassing technique to a variety of biological datasets.

Sambyal et al. [11] used sophisticated deep learning algorithms to predict disorders of the eyes, kidneys, and cardiovascular system in people with diabetes. Their evaluation technique included rigorous statistical methods, with a particular emphasis on the widely used PIMA Indian Diabetes Dataset obtained from the UCI repository. This technique sheds light on the model's efficacy in predicting illnesses connected with diabetic complications.

In another significant study [12], the authors led the construction and assessment of multiple deep (CNNs) used to discern between normal and pathological frontal chest radiographs. The overall purpose of this program is to help radiologists and clinicians discover possible problems by providing a useful tool for prioritizing and triaging task lists and reports.

In one study [13] Nasser and team suggest a new way to figure out chest diseases They do this in two steps First they classify X ray images into three groups normal lung disease and heart disease In the second step they look at seven specific lung and heart diseases They use two methods one is a mix of deep learning models and the other is called VTChestNet performed the best beating other models.

In another study [14] Ms Kavitha and team check how well models can classify diseases and they explain the results using accuracy To make the models more accurate they do some things with the images before using them They use techniques to remove unnecessary noise from X ray images and make them clearer The dataset they use has a CSV file and a folder with X ray images showing six types of diseases totaling 1 120 X rays The study talks about CNNs as a tool for diagnosing chest problems explaining how they are built and how they work

III. Proposed Methodology

In this section, we provide a succinct overview of the proposed BGOPDP (BGO based Pulmonary Disease Prediction) model, as depicted in Figure 1, illustrating the sequential steps of the COVID-19 prediction model. To enhance understanding, a detailed explanation of each block, along with a set of equations, is presented in this section. Additionally, Table 1 elucidates the notations used in the proposed work and their respective meanings.

Input X-Ray Pre-Processing:

The initial step involves pre-processing Input X-Ray XI images to facilitate feature extraction. Adjustments are made to the dimensions of the input image to align with the working environment requirements. Subsequently, the image undergoes Wiener filter filtering to eliminate undesirable noise.

Filter: The primary role of the Wiener filter is to eliminate additive noise and minimize mean square error, acting as a linear estimate of the original image. This pre-processing step plays a crucial role in enhancing and refining the input data for the subsequent stages in the BGOPDP model. Hence mean(μ) and variance (σ^2) of each pixel are estimated by Eq. i, ii and Eq. iii [13].

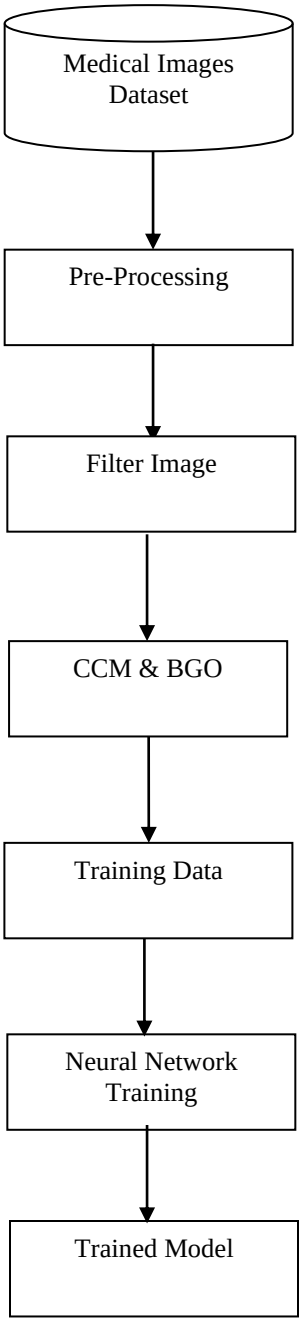


Fig. 1 Block diagram of proposed model**.**

$$\mu_{x,y} = \frac{1}{XY}\sum_{m,n\in N} XI(m,n) \quad \text{----------1}$$

$$\sigma^2_{x,y} = \frac{1}{XY}\sum_{m,n\in N} XI(m,n)^2 - \mu^2 \text{----------2}$$

Where XI is an input image, having filter, NxM N=M=3, x and y are positioned in NxM block.

$$XIP(x,y) = \mu + \frac{\sigma^2 - v^2}{\sigma^2}\left(XI(x,y) - \mu\right) \text{ ----------3}$$

Changes as per Weiner filter is done by Eq. 3 where v^2 is noise variance.

CCM Feature Extraction

A method employed to capture the surface characteristics of an image is through the use of a co-occurrence matrix, serving a specific purpose. The co-occurrence matrix's surface attributes are manifested through the interrelation of adjacent pixels [15]. It quantifies and characterizes the surface components, utilizing four key elements in this study: contrast, energy, inverse difference, and entropy.

Biogeographically Optimization

In this step image is divide into two part featured portion uses for the training and other is non featured part.

Habitat Suitability Index (HSI): This represents the fitness value of a habitat; a higher value signifies an unfavorable location for habitation, while a lower value indicates a favorable area in terms of resources, life, and other factors.

Immigration and Emigration Rate: Basic terms for immigration (λ) and emigration (α) are determined by equations 8 and 9 [16].

$$\lambda_R = (1 - \alpha_R) \text{----------8}$$

$$\alpha_R = \frac{R}{h} \text{----------Eq. 9}$$

Where R is rank of habitat in terms of HSI value, while h is total number of habitats.

Generate Habitats: In this phase, a set of potential solutions is created, termed as "habitats" in this method. Each habitat comprises probable cluster centers, making it a combination represented as H=U1, Us, where the population consists of a total of h habitats. Equation 10 outlines the population generation function in this approach.

$$H \leftarrow \text{Habitat}(s, h) \text{----10}$$

Fitness Function (HSI): The suitability of a habitat is determined by its index, calculated based on distance. It involves the summation of the shortest distances between the representative pixels of a segment and the pixels segmenting the image. In the current population, chromosomes with the lowest summation value are considered the optimal solution.

$$HFV \leftarrow \text{SUM}(\min(H_{h,s}, XIP)) \text{ ----------11}$$

In Eq 10 F_h denotes the fitness value of the h^{th} habitate in the population H and $H_{h,s}$ represents a typical pixel from the h^{th} habitate and s^{th} segment As a consequence the rank i of the habitat is derived by adding the F_c value distances between each habitat and the other user

$$R = \text{Rank}(F_h, H) \text{----12}$$

Crossover

The emigration rate dictates the pace at which users transition from one environment to another, while the immigration rate governs the admission of species into a habitat. Consequently, the determination of crossover from one habitat to another requires the discovery of both emigration and immigration rates. Thus, the crossover is contingent on the fulfillment of the following condition [17].

Mutation

Following crossover, mutation is also implemented in this study to increase the likelihood of obtaining novel solutions. The mutation probability is introduced, wherein mutation is applied to selected habitats based on the Habitat Suitability Index (HSI) value.

$$M_h = \frac{R_h}{sum(h)} \text{----------13}$$

$$M_p = \frac{M_h}{Max(M_h)} \text{ ----------14}$$

Therefore, for habitats that cross a constant mutation_cross_limit within the range of 0-1, M_h provides a mutation rank for the habitat based on the HSI value. A higher value indicates a higher mutation rank. Consequently, habitats with a higher mutation rank exhibit a higher mutation probability. Habitats with a

mutation probability lower than the Mutation-Cross Threshold undergo mutation.

Training of Neural Network

In the context of neural networks, a theoretical framework is considered with the assumption that it comprises three layers. These layers consist of neurons, where the neurons in the input layer are denoted by 'I,' those in the hidden layer are denoted by 'j,' and the neurons in the output layer are represented by 'k.'

The connections between these neurons are characterized by weights, denoted by 'wij,' where 'i' and 'j' correspond to the respective neuron layers. The output of a neuron, as per the weighted sum and a biasing value 'bj,' is expressed by Equation 11. This equation serves as a representation of the computational process within the neural network, highlighting the role of weights and biases in determining the output of each neuron.

$$X_j = \sum x_i \, . \, w_{ij} \, b_j \qquad 15$$

where, $1 \leq i \leq n$, n is the number of inputs to node j, and bj denotes the biassing for node j. As a result, the network will learn the weights between layers [20]. The weight values of each layer must be adjusted to remedy this issue. So eq. 15 [18] was used to estimate error.

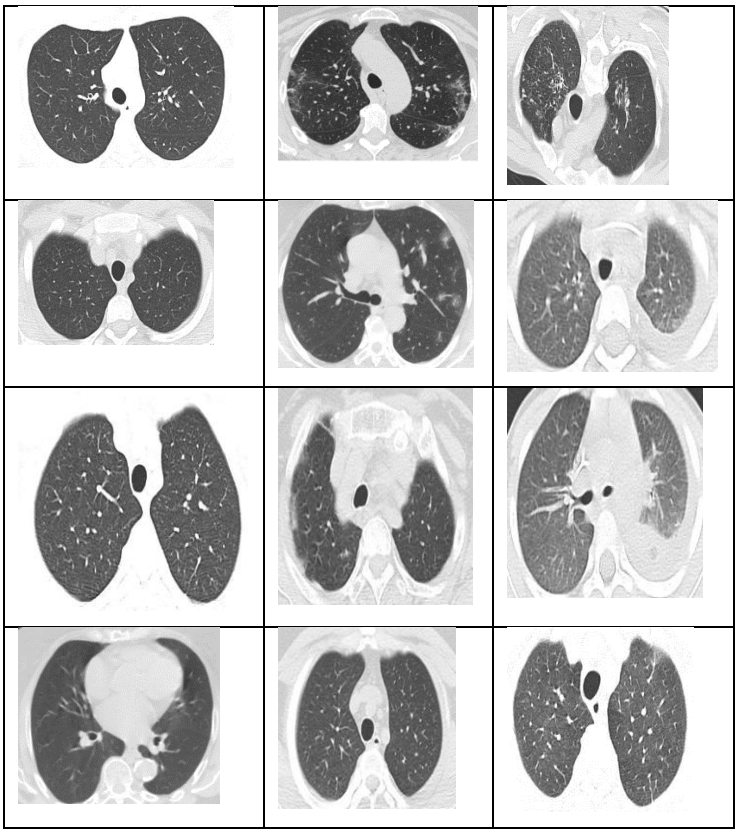
IV. Experiment and Results

The classification of X-ray chest images was implemented using MATLAB. The experimental machine was equipped with an I3 6th generation processor and 8GB RAM. The proposed model was compared with existing models suggested in [19] and SIFCD [20]. The experimental dataset utilized in the study was obtained from [21], and a comprehensive description of the dataset is provided in Table 2.

Dataset

Table 1 Multi Class dataset images.

Healthy	Covid	Other Pulmonary



Unprecedented Scale and Diversity

Dataset is unparalleled in its scale and diversity. It comprises 4,173 CT scans from 210 different patients, each meticulously collected from two renowned healthcare institutions.

Healthy Patients: We have included 758 CT scans from healthy individuals, with an average of 15 scans per patient. These scans serve as a vital reference point for distinguishing between normal lung conditions and those affected by SARS-CoV-2.

SARS-CoV-2 Infected Patients: For a comprehensive analysis of COVID-19, our dataset includes 2,168 CT scans from 80 patients who were infected with SARS-CoV-2 and confirmed through reverse transcription polymerase chain reaction (RT-PCR). These patients underwent an average of 27 CT scans each during the course of their illness, allowing for a detailed progression study.

Other Pulmonary Conditions: To enhance the dataset's utility and versatility, we've incorporated 1,247 CT scans from patients diagnosed with various pulmonary conditions unrelated to COVID-19, averaging 16 scans per patient.

This inclusion aids in distinguishing between COVID-19 and other lung disorders, enabling more accurate diagnoses.

Results

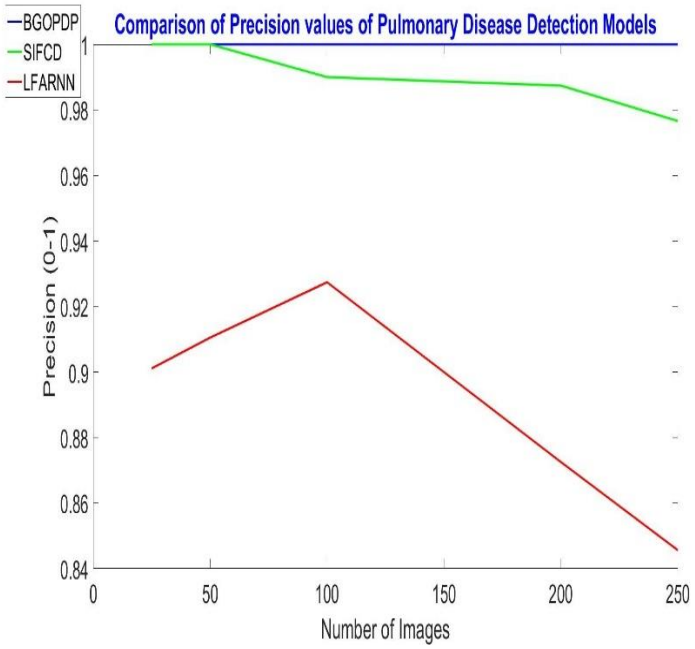


Fig. 2 Comparison of precision values for pulmonary disease detection models.

The precision values for various pulmonary disease detection models are visually represented in Figure 2. Notably, upon analysis, it was observed that the proposed model consistently exhibits the highest precision values across all sets of experimental images. These experimental sets specifically pertain to a binary testing scenario, where the input images undergo classification into either the infected or healthy class. The discernible trend in precision values suggests the superior accuracy and reliability of the proposed model in distinguishing between different pulmonary conditions. This finding underscores the effectiveness of the model in binary classification tasks, emphasizing its potential as a robust tool for precise and dependable pulmonary disease detection.

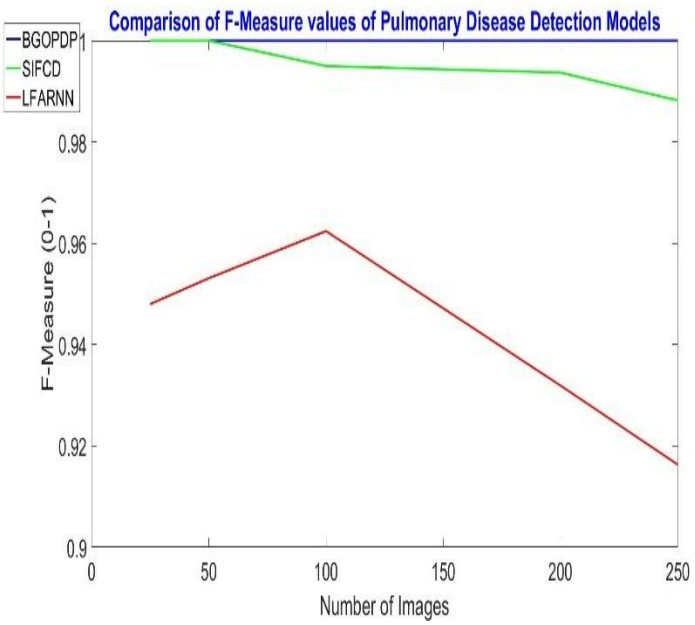


Fig. 3 Comparison of f-measure values for pulmonary disease detection models.

F-measure value of different pulmonary disease detection models is shown in fig. 3. It was found that proposed model has highest values in all set of experimental images. These are for the binary set of testing where input image was classified into infected or healthy class.

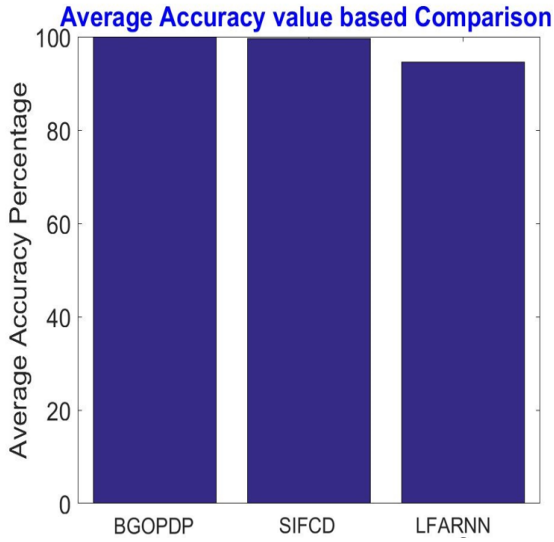


Fig. 4 Comparison of average accuracy values for pulmonary disease detection.

Fig. 4 shows the average accuracy comparison of all set of experimental pulmonary images. Result shows that BGOPDP has increased the detection accuracy.

Table 2 Multi class pulmonary disease detection Accuracy value based image comparison.

Types of Images	BGOPDP	SIFCD	Previous Model
Covid	**100**	**58.2**	**97.95**
Healthy	**100**	**99.98**	**78.14**
Other Pulmonary	**99.97**	**100**	**46.76**

Table 3 Multi class pulmonary disease detection error value based image comparison.

Types of Images	BGOPDP	SIFCD	Previous Model
Covid	0	41.8	2.05
Healthy	0	0.02	21.86
Other Pulmonary	0.03	0	53.24

The values presented in Table 2 and Table 3 depict the outcomes of chest pulmonary multiclass image classification models. Notably, the proposed BGOPDP model exhibits improvements over its predecessor, demonstrating enhanced performance. The incorporation of a neural network for learning contributes significantly to the efficacy of the model. Furthermore, the utilization of histogram features and genetic cluster features serves to augment the learning process of the model. The integration of these features enhances the model's ability to discern and classify images accurately. A notable achievement is observed in the form of a 13.93% improvement in accuracy attributed to the BGOPDP model when compared to the SIFCD model [23]. This progress underscores the effectiveness of the proposed model in advancing the state-of-the-art in chest pulmonary multiclass image classification, emphasizing the role of innovative features and neural network-based learning mechanisms in achieving superior results.

V. Conclusions

This paper introduces a model designed for predicting pulmonary diseases in medical images. The content feature co-occurrence matrix is extracted from the images. To enhance the learning capabilities of the model, the proposed approach involves clustering the image into two regions, thereby reducing calculation costs. The model is capable of handling both two-class and multiclass predictions for pulmonary diseases. Experiments were conducted on real datasets, demonstrating the effectiveness of the proposed model across all datasets. The BGOPDP model outperforms SIFCD, showcasing a 0.92% increase in precision. For multiclass prediction in pulmonary datasets, BGOPDP exhibits a notable improvement of 13.93% compared to SIFCD [20]. Future research endeavors may explore the application of this model to various types of pulmonary medical images for disease detection and prediction.

References

[1]. T. Sahlol, M. Abd Elaziz, A. Tariq Jamal, R. Damaševičius, and O. Farouk Hassan, “A novel method for detection of tuberculosis in chest radiographs using artificial ecosystem-based optimisation of deep neural network features,” Symmetry, vol. 12, no. 7, p. 1146, 2020.

[2]. M. A. Khan, V. Rajinikanth, S. C. Satapathy et al., “VGG19 network assisted joint segmentation and classification of lung nodules in CT images,” Diagnostics, vol. 11, no. 12, p. 2208, 2021.

[3]. CPoap, M. Wozniak, R. Damaševičius, and W. Wei, “Chest radiographs segmentation by the use of nature-inspired algorithm for lung disease detection,” in Proceedings of the 2018 IEEE Symposium Series on

Computational Intelligence (SSCI), pp. 2298–2303, IEEE, Bangalore, India, November 2018.

[4]. J. Suri, S. Agarwal, P. Elavarthi et al., “Inter-variability study of COVLIAS 1.0: hybrid deep learning models for COVID-19 lung segmentation in computed tomography,” Diagnostics, vol. 11, no. 11, p. 2025, 2021.

[5]. Rahman, “CXR data set repository,” 2020, https://bit.ly/3p6KtQD Kaggle.

[6]. T. Rahman, A. Khandakar, Y. Qiblawey et al., “Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images,” Computers in Biology and Medicine, vol. 132, Article ID 104319, 2021.

[7]. M. E. H. Chowdhury, T. Rahman, A. Khandakar et al., “Can AI help in screening viral and COVID-19 pneumonia?” IEEE Access, vol. 8, pp. 132665–132676, 2020.

[8]. Gabruseva, T.; Poplavskiy, D.; Kalinin, A. Deep learning for automatic pneumonia detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Virtual Conference, 14–19 May 2020; pp. 350–351.

[9]. Ibrahim, A.U.; Ozsoz, M.; Serte, S.; Al-Turjman, F.; Yakoi, P.S. Pneumonia classification using deep learning from chest X-ray images during COVID-19. Cogn. Comput. 2021, 1–13.

[10]. Hasan, S.; Shamsuddin, S.M. Multi-Strategy Learning and Deep Harmony Memory Improvisation for Self-Organizing Neurons. Soft Comput. 2019, 23, 285–303.

[11]. Sambyal, N.; Saini, P.; Syal, R. A Review of Statistical and Machine Learning Techniques for Microvascular Complications in Type 2 Diabetes. Curr. Diabetes Rev. 2021.

[12]. Tang, YX., Tang, YB., Peng, Y. et al. Automated abnormality classification of chest radiographs using deep convolutional neural networks. npj Digit. Med. 3, 70 (2020).

[13]. Nasser, A.A., Akhloufi, M.A. Deep Learning Methods for Chest Disease Detection Using Radiography Images. SN COMPUT. SCI. 4, 388 (2023).

[14]. S. Ms. Kavitha, S. Thaarani, A. P. Singh and G. Santhosh, "Chest Disease Classification Using Convolutional Neural Networks," 2023 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2023.

[15]. Shalinee Jain, Asst. Prof. Sachin Malviya. "Digital Image Retrieval Using Annotation, CCM and Histogram Features". IJSRET, volume 4 issue 5, 2018.

[16]. Mohammed Alweshah. "Construction biogeography-based optimization algorithm for solving classification problems". Neural Computing and Applications, Springer volume 28 February 2018

[17]. MacArthur R., Wilson E. *The Theory of Biogeography*. Princeton, NJ, USA: Princeton University Press; 1967.

[18]. K. El Asnaoui and Y. Chawki, “Using X-ray images and deep learning for automated detection of coronavirus disease,” J. Biomol. Struct. Dyn., vol. 39, no. 10, pp. 3615–3626, 2021.

[19]. Chang Min Kim, Ellen J. Hong, Roy C. Park. "Chest X-Ray Outlier Detection Model Using Dimension Reduction and Edge Detection". IEEE Access, June 22, 2021.

[20]. Priyanka Singh, Dr S. Veenadhari. "Chest Image Based COVID Prediction Using Genetic Cluster and Histogram Feature". 1303-5150, 2022.

[21].]https://github.com/UCSD-AI4H/COVID-CT/tree/master/Images-processed.